Computing Interdisciplinarity of Scholarly Objects using an Author-Citation-Text Model

Min-Gwan Seo, Seokwoo Jung, Kyung-min Kim, and Sung-Hyon Myaeng

Korea Advanced Institute of Science and Technology,
291 Daehak-ro (373-1 Guseong-dong), Yuseong-gu, Daejeon 305-701, South Korea,
smksyj@kaist.ac.kr, wjdtjrdn1201@kaist.ac.kr, kimdarwin@kaist.ac.kr,
myaeng@kaist.ac.kr

Abstract. There has been a growing need to determine if research proposals and results are truly interdisciplinary or to analyze research trends by analyzing research papers, reports, proposals and even researchers. In this paper, we tackle the problem and propose a method for measuring interdisciplinarity of scholarly objects. The newly proposed model takes into account authors, citations, and text content of scholarly objects together by building author networks, citation networks and text models. The three types of information are mixed by building network embeddings and sentence embeddings, which rely on the network topology and context-driven word semantics, respectively, through neural network learning. In addition, we propose a new measure that considers not only evenness of disciplines but also distributions of the magnitudes of disciplines so that saliency of disciplines is well represented.

Keywords: Interdisciplinary research, Document classification, Network embedding, Document embedding, Scientometrics

1 Introduction

There has been a growing need to determine if research proposals and reports are truly interdisciplinary. For example, when government funding agencies and universities want to promote interdisciplinary research, they need to search for re-search proposals and reports for the degree of interdisciplinarity [2, 3, 11, 12]. In an effort to address this issue, some past studies sought to apply new measures borrowed from ecology and economics [1, 12], but without modeling documents for their contents and inter-document relations.

In this paper, we tackle the problem and propose a method for measuring an interdisciplinarity (ID) score of scholarly objects such as individual documents, journals and conference proceedings. The newly proposed model takes into account authors, citations, and text content of scholarly objects together to determine the disciplines they represent. A distribution of disciplines is computed by building an author network, a citation network and text models. In addition, we propose a new measure that considers not only the number of disciplines but their magnitudes so that saliency of disciplines can be considered.

The overall process of calculating an ID score for interdisciplinarity of a scholarly object is carried out as follows. The input article, for example, is analyzed not only for its textual content but also for its citations and authors so that they are projected onto the corresponding networks built for the entire collections. Essentially three vectors are constructed for the three aspects of the scholarly object and combined to form its feature vector, which is fed into a classifier that determines its subject categories. The result is a distribution of subject category (or discipline) strengths. The next step is to compute an ID score.

2 Related Work

2.1 Detecting Disciplines of Scholarly Objects

The way scholarly object is viewed and analyzed is a key to determining interdisciplinarity of a scholarly object. However, most of the previous studies simply used citation counts to get a distribution of discipline strengths. The work in [11, 5] calculated interdisciplinarity for journals in Web of Science (WoS). They assigned journals to 244 subject categories defined by WoS and then grouped the subject categories into 6 macro-disciplines. To obtain a distribution of disciplines for a journal, they used article citation information in journals. In order to identify the names of journal embedded in the reference text, they applied simple text processing tools. [1] proposed a method focusing on interdisciplinarity of research teams. To compute a distribution of disciplines for a team, they counted the disciplines to which the researchers in the team belong. They used all the categories/disciplines of the PhD groups and 205 department groups. In order to analyze interdisciplinarity of target journals, the work in [9] used various types of bibliometric indicators and 225 subject categories defined by ISI in its SciSearch database.

2.2 Interdisciplinarity Measures

Previous studies borrowed the concept of interdiciplinarity from the diverse sources like ecology and economics [1, 12, 14] where diversity is defined with three different attributes:

- variety: the number of distinct categories (disciplines)
- balance: the evenness of the distributions of distinct categories
- disparity: the degree to which the categories are different

Table 1 shows interdisciplinarity measures defined previously. Any two categories (or disciplines) i and j are represented with their proportions p_i and p_j in the system, while $d_{i,j}$ and $s_{i,j}$ measure the distance and the similarity between the two. The category count measure uses only a single attribute, but others consider multiple attributes. The Shannon entropy, Simpson index, and Stirlings diversity measures are most frequently used in computing interdisciplinarity in the literature. To the best of our knowledge, however, no studies have compared the different measures for their relative merits with real life data.

Name	Attribute	Form
Category count	Variety	N
Shannon Entropy	Variety / Balance	$-\sum_{i=1}^{N} p_i \log p_i$
Simpson	Variety / Balance	$\sum_{i=1}^{N} p_i^2$
Total Dissimilarity	Disparity	$\sum_{i,j}(1 - s_{i,j})$
Stirling's diversity	Variety / Balance / Disparity	$\sum_{i,j} d_{i,j} p_i p_j$

Table 1: Interdisciplinarity measures defined with three attributes of diversity

3 Proposed Methods

3.1 Author-Citation-Text Model

The **ACT (Author-Citation-Text) model** combines information obtainable from three sources to compute a distribution of disciplines for a scholarly object: the author network that captures co-authorships among the articles in the collection, the citation network that connects articles based on direct citations, and the text (abstracts in this work) of the articles. In short, an article is represented jointly with the local text and its relative positions in the global networks (i.e. author and citation networks).

Citation information is important in capturing the extent to which the article cites others under different disciplines. The more articles under different disciplines are cited, the more interdisciplinary the target article would be. However, citation information alone may not be sufficient because the citation network is constructed based on explicit citations. For example, there may not be an explicit citation for text content under a different discipline if it is too old to cite or sufficiently well known to the community. In order to compensate for this limitation, we consider co-authorship relationships and the semantics of text. The latter is important because different disciplines would have developed their own vocabularies, which can serve as discriminative features for a classifier.

In the proposed model, the three different types of information for articles are used to represent each article as a vector comprised of three corresponding vectors. An article vector then becomes an input to a research category classifier which returns a vector whose elements correspond to available research categories (19 in the current system), which is also referred to as a distribution of disciplines. We built two classifiers: one based on a neural network with two hidden layers and the other based on logistic regression.

Vectors corresponding to author, citation and text information are generated with embedding methods. Given a citation network connecting all the articles, where a node and an edge represent an article and a citation relation, respectively, we generate a vector for each article by employing a network embedding method that considers node connectivity information in a graph structure [10]. We employ the same method for an author network where a node and an edge represent an author and a co-authorship relation, respectively. A piece of text (an abstract in the current implementation) is also represented as a vector by

Algorithm 1 Proposed Interdisciplinarity Measure

1: **procedure** IDScore(D)
2: **input:** distribution of disciplines D
3: **output:** interdisciplinarity score based on the salient discipline set
4: $H, L \leftarrow partition(D)$
5: **return** $\frac{\sum_i (p_i + L_1) log(p_i + L_1)}{p_1} |H| \sum_{i,j} d_{i,j} \qquad \forall_{i,j} \in H$

employing a document embedding method [4], which is an extension of a word embedding method [8].

After the document/network embedding step, we obtain a vector $\mathbf{v}$ for an article, which is formed by concatenating the author vector $\mathbf{a}$ of dimension 64, the citation vector $\mathbf{c}$ of dimention 64, and the text vector $\mathbf{t}$ of dimension 300 for the article. A text vector of dimension 300 is trained by the doc2vec algorithm. An author vector is obtained by averaging the vectors for the authors who wrote the same article. Likewise, a citation vector is constructed in the same way. The network vectors are trained in advance by the *deepwalk* algorithm [10].

After obtaining an article vector $\mathbf{v}$, we can now compute the distribution of disciplines for the article through a classifier. While any classifier would work for this purpose, we employed a neural network classifier and a logistic regression classifier. The classifier is pre-trained to predict the distribution of a given article vector $\mathbf{v}$.

3.2 Proposed Interdisciplinarity Measure

Given a distribution of disciplines for an article or any scholarly object, we compute an ID score that captures the degree to which it is interdisciplinary. While Stirlings diversity introduced in Table 1 has been used widely used in previous studies of interdisciplinarity [6, 12], we propose a new measure that is not overly biased toward a large number of disciplines involved in a scholarly object.

Let us consider an article in bioinformatics, which is clearly an interdisciplinary area between biology and computer science. Even if the article just covers the two disciplines, it should not be considered less interdisciplinary than an article covering more than two areas just because of the number of areas. This is an indication that the number of disciplines is not a critical factor for interdisciplinarity as long as it contains at least two salient disciplines.

Based on this observation, the proposed measure attempts not to value the number of disciplines or consider all the expressed disciplines in calculating interdisciplinarity. Instead, it focuses on salient disciplines with high proportions in the distribution of disciplines. As a way of identifying salient disciplines, we introduce a partitioning method that divides disciplines based on their magnitudes in the distribution. The partition function is described in Algorithm 1 that shows the overall steps for computing an ID score. It selects salient disciplines based on one of the following two methods: *largest gap* and *k-means with initialization*.

The *largest gap* method sorts the disciplines in a distribution by an descending order of their magnitudes and then calculates the differences between two neighboring disciplines. Disciplines are partitioned into two by drawing a line between the two disciplines showing the highest difference. The **k-means with init** methods clusters the disciplines into two groups (i.e. k=2) based on their magnitudes as their unique attributes so that we can separate the disciplines with high magnitudes (or salient ones) from those with small magnitudes (or not-so-salient ones). To prevent the randomness of k-means, we set the two initial points with the biggest and smallest values.

After the partitioning, the following formula is used to calculate the ID score from the salient disciplines.

$$\frac{\sum_i (p_i + L_1) log(p_i + L_1)}{p_1} |H| \sum_{i,j} d_{i,j} \qquad \forall_{i,j} \in H$$

where H and L correspond to salient and not-so-salient partitions, respectively. In the formula, p_i is a proportion of the i-th discipline in a salient partition, and therefore p_1 is the biggest proportion in the salient partition. Likewise, L_1 is the biggest proportion in the not-so-salient partition. The distance $d_{i,j}$ between the disciplines i and j is consider for every pair. Various types of distance measures such as Euclidean distance and Cosine distance can be used for the distance term.

This formula includes three key factors. First, we consider the size of the salient disciplines ($|H|$) for diversity. Second, we use the distance among the salient disciplines, which is computed with the sum of distances between all disciplines in the salient set ($d_{i,j}$). Third, the *log* term is a modified entropy that focuses on the degree of integration among salient disciplines where L_1 is the biggest value from other discipline set. This is used to add the information of the not-so-salient disciplines. This modified entropy is divided by the biggest value from the salient discipline set (p_1) for normalization.

Basically, its *log* term is based on the entropy within the salient discipline set; this score increases when the distribution of salient disciplines are even. And because of the distance term ($d_{i,j}$), this score increases when the disciplines in the salient set are different. The score also increases if there are multiple disciplines considered as salient because of the size term ($|H|$).

4 Evaluations

4.1 Dataset

We used bibliographic data of research articles from Microsoft Academic Search (*MAS*) which includes meta-data such as titles, authors, fields of study, references, keywords, and abstracts. From this meta-data, *fields of study* contains keywords that show the subject fields of the paper, such as 'physics' or techniques such as 'logistic regression' and 'wireless network.' The values under *keywords* are automatically extracted from the title and abstract of a given paper. We

6

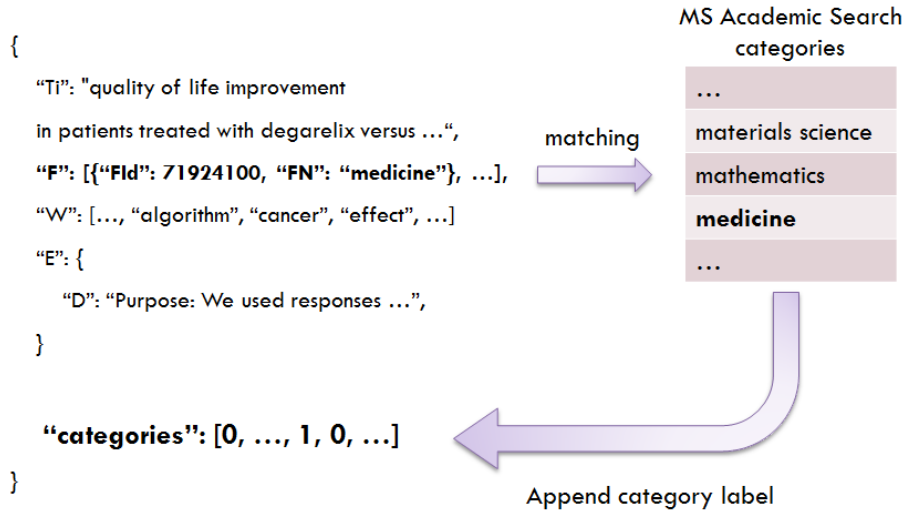

Fig. 1: Paper labeling Flow

collected data from 1994 to 2015 for every three years, mainly due to the lack of storage and computing time. For more qualitative analysis involving human judgments, we only used 2015 data as older articles tend to have more missing meta-data.

To construct a gold standard for training and testing, we assign the 19 discipline fields to all the articles. That is, the task of the proposed method as well as some baselines is to assign discipline labels as close to the gold standard as possible. Because the articles in *MAS* do not have discipline field labels, we devised a method for automatically assigning labels. This discipline field labeling method compares those of *fields of study* in each article against the 19 major discipline labels and 268 sub-discipline labels pre-defined by *MAS*. For the comparison, we used exact string matching between each value under *fields of study* and each discipline. Figure 1 shows the overall process for automatic paper labeling.

Pre-defined major discipline labels are as follows: **Art, Biology, Business, Chemistry, Computer Science, Economics, Engineering, Environmental Science, Geography, Geology, History, Materials Science, Mathematics, Medicine, Philosophy, Physics, Political Science, Psychology, Sociology**

Because the *keywords* part includes automatically extracted words from the paper, using them for labeling can assign irrelevant labels to the paper. Therefore, we decide to use the values of *fields of study* for conservative labeling. The matching discipline name or the parent of the matching sub-discipline name was assigned to the article. After the labeling step, an article can contain multiple discipline labels if its fields of study have multiple names.

Table 2 shows the number of unique authors and articles. From the table, the *whole articles* are collected from *MAS*. The *target articles* are obtained after

Year	1994	1997	2000	2003
# of unique authors	2,345,006	2,459,321	2,561,151	2,809,515
# of target articles	141,367	208,253	359,657	487,069
# of whole articles	1,697,000	1,748,000	1,769,000	1,921,000
Year	2006	2009	2012	2015
# of unique authors	3,026,501	3,218,836	3,453,158	3,608,795
# of target articles	681,916	800,585	786,753	593,515
# of whole articles	1,818,000	1,698,000	1,663,000	1,648,000

Table 2: The number of unique authors and articles

eliminating the articles without citation and/or abstract fields from the whole articles.

4.2 Evaluation in Precision for Individual Articles

In order to evaluate the proposed method, we used author network, citation network and text embeddings as features for logistic regression and neural network classifiers. The proposed ACT model was compared against its sub-components, namely citation (C) only and a combination of citation and author information (AC), which have been used previously. We chose to use a logistic regression and neural network classifiers because the former is well-known for general effectiveness and the later for its popularity based on high performance. For the neural network classifier, we used two hidden layers with 256 and 128 nodes. Following well-known parameter settings, we employed Rectified Linear Unit (ReLU) for the activation function and Categorical cross-entropy for the loss function. AdaDelta was used as an optimizer, and a simple SGD with 128 mini-batch was used for gradient updates.

For the logistic regression classifier, we used cross-entropy for a loss function and L2 penalty for regularization. This classifier was trained under the one-vs-rest scheme for the multi-label problem. For all the learning and testing, we used 10-fold cross validation with all the collected articles. For each cross validation step, we train classifiers by using author, citation, text information and labels from papers in the training set and predict the distribution of papers in the test set.

Because the output of the classifier is a distribution of the disciplines whereas each of the article labels are treated as a binary value, both label-oriented and distribution-oriented evaluations were conducted. For the former, we define label precision (LP) as follows:

$$LP(X,Y) = \frac{\sum_{i=0}^{|Y|} \mathbb{1}_Y(x_{i,dis})}{|Y|}$$

where X is a set of disciplines sorted in a descending order of the magnitude and Y is a set of disciplines whose label is 1. $\mathbb{1}_Y(x_i)$ is an indicator function that returns 1 iff $x_i \in Y$ otherwise returns 0. $x_{i,dis}$ is the discipline of i-th element

Classifier Type	Logistic Regression			Neural Network		
Feature Combinations	C	AC	ACT	C	AC	ACT
Average Label Precision	0.8103	0.8097	0.8478	0.8332	0.8250	0.8497
Average JS-divergence	0.2852	0.2853	0.2525	0.1532	0.1560	0.1313

Table 3: Label Precision and JS-divergence results for different feature combinations and classifiers

in X. Label precision measures the extent to which the automatically assigned labels are true discipline labels are highly ranked in the predicted discipline distribution. For the distribution-oriented evaluation, we use Jansen-Shannon divergence (JSD) [7], which measures dissimilarity between two distributions. Hence, the lower JSD, the better the predicted discipline distribution.

Table3 shows average LP and JSD results over all the evaluated articles. ACT shows the best performance in both of the classifiers, confirming the superiority of the proposed method compared to the baseline of using citations only. It should be noted that using citation information alone through the network embedding, i.e. making use of the disciplines of the cited articles, already gives a relatively high performance (0.8332% in LP and 0.1532 in JSD with the neural network classifier).

It is also worth noting, however, that when the author information is added to the citation information (AC), the performance decreases slightly in both LP and JSD. Our analysis indicates that the author embeddings were not as effective as citation embeddings because the author network built on the co-authorship relations is much sparser than the citation network. With a small network and low weights on edges caused by small co-authorship frequencies, the training through author embeddings gave an adversary effect to the final classification. Left for future research is a comparison against the use of author affiliations like departments in a direct way, although it may have a bias because an inclusion of an author from a different disciplinary department may not make a research so interdisciplinary.

4.3 Evaluation under Different Interdisciplinarity Measures

This part of evaluation has dual goals: one is to validate the proposed measure against the previously used ones in determining interdisciplinarity and the other is to evaluate the proposed model under the different measures of interdisciplinarity. For more qualitative analysis, we constructed a ground truth based on human judgments of journals and conferences. We first randomly selected 100 journals/conferences published in 2015 from a total of 1156 journals/conferences that contain more than 100 papers. The goal is to ensure that the selected journals/conferences should have enough data for the proposed ACT model.

Six human raters were asked to evaluate the interdisciplinarity of each journal/conference with scores ranging from 1 to 5 based on its introduction, aims, and research interest pages. The raters were given a guideline that specifies the

	Logistic Regression							
Feature	C				ACT			
Interdisciplinarity Measure	Entropy	Stirling	Salient (lg)	Salient (km)	Entropy	Stirling	Salient (lg)	Salient (km)
Spearman Correlation	0.3272*	0.3847*	0.4101*	0.4877*	0.3689*	0.4396*	0.3686*	0.5401*
	Neural Network							
Feature	C				ACT			
Interdisciplinarity Measure	Entropy	Stirling	Salient (lg)	Salient (km)	Entropy	Stirling	Salient (lg)	Salient (km)
Spearman Correlation	0.4960*	0.5527*	0.4422	0.4299	0.5071*	0.5700*	0.4826*	0.5732*

* Correlation is statistically significant at the 0.05 level

Table 4: Spearman Correlation between human ratings and those with different interdisciplinarity measures under different feature vectors and classifiers

number of disciplines covered, the evenness of the included disciplines, and their differences for different ratings. The journals/conferences that did not obtain four or more votes for a particular score were excluded to ensure credibility of the test collection. As a result, only 75 journals/conferences remained with their scores for interdisciplinarity.

This ground truth data was used to evaluate different interdisciplinarity measures including the proposed one, under which the proposed model is also evaluated to see its superiority from different perspectives. We adopted the Spearmans rank correlation coefficient [13] between the human judges and computed scores to compare among the interdisciplinarity measures. For an interdisciplianrity value of a journal/conference computed with a particular measure, we took the mean of the values computed for all the articles in it.

In order to compute the distance between two disciplines $d_{i,j}$ for the Stirling's measure and Salient measures, we created a discipline-discipline citation matrix X where $X_{i,j}$ is the citation count from discipline i to j so that we calculate $d_{i,j} = 1 - cos(X_i, X_j)$ where X_i is a citation vector from i th discipline to all other disciplines and $cos(X_i, X_j)$ is the cosine similarity.

Table 4 shows the Spearman correlation between the ground truth and the rankings under the interdisciplinarity measures for the cases of using C and ACT vectors and the two classifiers. *Salient(lg)* and *Salient(km)* mean the proposed interdisciplinarity measure with the *largest gap* and *k-means with initialization* methods, respectively. We show the results for only two different feature vector types because AC was known to be problematic in the previous experiment.

Most notable in the result is that regardless of the classifiers or the models, the proposed measure is most similar to the ground truth or the way human judges evaluate interdisciplinarity of the journals/conferences. While the guideline was geared toward the intended evaluation aspects included in the newly proposed measure, this result confirms that the new measure follows more closely the human raters qualitative judgements. It is also worth noting that the Stirlings

measure is consistently superior to entropy, perhaps because of its consideration of diversity or the distance between disciplines.

The result also confirms the superiority of the proposed model under the different measures including the new one. The highest correlation 0.5732 was obtained with the ACT features and the neural network classifier. Between the two different ways of selecting salient disciplines, the one with k-means clustering was consistently superior.

5 Conclusions and Future Work

This paper addresses the problem of computing the degree of interdisciplinarity of a scholarly object. We identified two problems: one is the lack of a proper model for determining the disciplines represented by a scholarly object, and the other is the way interdisciplinarity is measured. For the first one, we propose the Author-Citation-Text joint model that predicts the distribution of disciplines in a scholarly object based on the learned citation and author embeddings and document embeddings. For the second problem, we propose a new measure that takes into account saliency of disciplines appearing in a scholarly object.

From the experiment with a collection of articles over multiple years, the proposed model shows that the combination of the three aspects of articles can predict the discipline distributions more accurately. We also conducted a separate experiment for a more qualitative analysis by constructing a gold standard of 75 journals/conferences based on human judgments. Comparing the Spearmans correlation between human judgments and the automatically computed interdisciplinarity shows that the proposed measure captures the intended aspects of interdisciplinarity and that the proposed model is also superior under different measures.

The current work is novel in its tackling of the two key issues, modeling of scholarly objects for determining discipline distributions and measuring of interdisciplinarity. In addition, the construction of the two test collections for evaluations is also a significant contribution. However, we consider this work has several limitations. First, we did not fully explore different ways of using author information, other than just building an author network and author embeddings. Likewise, there is a plenty of room for considering different ways of combining citation and text information and even constructing different representations as the techniques for document and network embeddings are still in progress. Second, there is a room for improving the quality of the collections we developed for evaluations. While the way the discipline labels were attached to the individual articles is reasonable and quite accurate, it should be more carefully examined for accuracy perhaps by employing human experts or by means of crowdsourcing.

References

1. Aydinoglu, A.U., Allard, S., Mitchell, C.: Measuring diversity in disciplinary collaboration in research teams: An ecological perspective. Research Evaluation **25**(1), 18–36 (2016)

2. Huutoniemi, K., Klein, J.T., Bruun, H., Hukkinen, J.: Analyzing interdisciplinarity: Typology and indicators. Research Policy **39**(1), 79–88 (2010)
3. Laudel, G., Origgi, G.: Introduction to a special issue on the assessment of interdisciplinary research. Research Evaluation **15**(1), 2–4 (2006)
4. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML, vol. 14, pp. 1188–1196 (2014)
5. Leydesdorff, L., Rafols, I.: A global map of science based on the isi subject categories. Journal of the American Society for Information Science and Technology **60**(2), 348–362 (2009)
6. Leydesdorff, L., Rafols, I.: Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. Journal of Informetrics **5**(1), 87–100 (2011)
7. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information theory **37**(1), 145–151 (1991)
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119 (2013)
9. Morillo, F., Bordons, M., Gómez, I.: An approach to interdisciplinarity through bibliometric indicators. Scientometrics **51**(1), 203–222 (2001)
10. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: KDD, pp. 701–710 (2014)
11. Porter, A.L., Rafols, I.: Is science becoming more interdisciplinary? measuring and mapping six research fields over time. Scientometrics **81**(3), 719–745 (2009). DOI 10.1007/s11192-008-2197-2. URL http://dx.doi.org/10.1007/s11192-008-2197-2
12. Rafols, I., Meyer, M.: Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. Scientometrics **82**(2), 263–287 (2010)
13. Spearman, C.: The proof and measurement of association between two things. The American journal of psychology **15**(1), 72–101 (1904)
14. Stirling, A.: A general framework for analysing diversity in science, technology and society. Journal of the Royal Society Interface **4**(15), 707–719 (2007)